Strategic Games and Algorithmic Transparency

Ignacio N. Cofone[*] and Katherine J. Strandburg[**]

Abstract

We challenge the common assertion that disclosing details of algorithmic decision-making processes ordinarily provides decision subjects with opportunities to "game the system," leading to inaccurate or unfair results. We delineate the limited situations in which such gaming is likely and develop normative considerations to distinguish, within that sub-set of gameable situations, those where the risk of gaming justifies opacity and those where transparency should be mandated irrespective of gaming. We argue that gameable situations should be distinguished based on the accuracy of the algorithmic proxy and the types of errors it tends to make. Overall, we argue that the range of decision contexts in which algorithmic opacity is justified are much narrower than is normally assumed, and we thereby hope to advance the discussion on algorithmic transparency.

Table of Contents

---

[*] Assistant Professor, McGill University Faculty of Law. ignacio.cofone@mcgill.ca.

[**] Alfred B. Engelberg Professor of Law, NYU School of Law. katherine.strandburg@nyu.edu. We thank xxx for their helpful comments. The paper also benefited from comments received at the Imperfect Enforcement Conference at Yale Law School and the Faculty Workshop at NYU School of Law.

I. Introduction

Readers probably heard of the Amazon survey tool MTurk, but may not know the story that gave it its name. The original Mechanical Turk was a chess-playing robot that was built in the 1800s by Wolfgang von Kempelen, who presented it as a machine that could play chess against humans. This "advanced A.I." managed to beat most of its opponents, including Napoleon Bonaparte and Benjamin Franklin, for more than 80 years. Eventually, however, the Turk was unveiled as a hoax: a small chess master would hide inside of the machine, operating it and playing against its challengers.

Today's algorithms are real – and used in highly consequential decision-making processes to compute credit scores, predict our riskiness as parolees or parents, assess our likely performance as students or employees and the like. Decision-makers use algorithmic assessments for a variety of reasons, such as reducing decision-making costs, reducing errors, or mitigating bias. Like a chess game against the original Mechanical Turk, an interaction with decision-making algorithms may feel like a game against a mysterious mechanical opponent. Ultimately, however, the algorithms are tools of human decision-makers, who control their design, implementation and ultimate impact. Their autonomy, like that of the Mechanical Turk, is an illusion.

Like the opponents of the original Mechanical Turk, we are often kept in the dark about the activities of the humans behind the algorithmic curtain, whose contributions are intentionally obscured behind claims of trade secrecy or fears of "gaming the system."[1] This paper attempts to peek behind the curtain to question

[1] Jack Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO STATE LAW JOURNAL (2017); Zachary Lipton, *The Mythos of Model Interpretability*, 16 QUEUE - MACHINE

this purposeful masking. While there is a considerable current literature discussing the risks of "black box" algorithms and the social benefits of disclosure and transparency,[2] we come at the issue from the other side of the equation, interrogating the basic argument, often simply assumed in policy debates, that the risk of gaming means that secrecy provides social benefits. [3] This leaves a literature gap: When is the gaming cost of algorithmic transparency serious enough to outweigh its benefits?

Interactions between decision subjects and decision-making algorithms, like more familiar interactions between human subjects and decision-makers, are strategic. None of these players can be assumed to be acting in society's interest. Decision subjects may seek better outcomes for themselves by gaming, but decision subjects' incentives to "game the system" are only part of the story. In choosing whether and what to disclose, algorithm creators and users also act strategically in pursuit of their own private ends. [4] Decision-makers' choices about disclosure can

---

LEARNING 30 (2018); Ignacio Cofone, *Servers and Waiters: What Matters in the Law of A.I.*, 21 STANFORD TECHNOLOGY LAW REVIEW 167 (2018).

[2] See, e.g., Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions Essay*, 89 WASH. L. REV. 1–34 (2014); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Solon Barocas, *Understanding Inscrutability*, ALGORITHMS & EXPLANATIONS CONFERENCE PAPER (2018); Andrew Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, FORDHAM LAW REVIEW (2019); SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (1 edition ed. 2018); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018).

[3] Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017); ANDREW GUTHRIE FERGUSON, THE RISE OF BIG DATA POLICING: SURVEILLANCE, RACE, AND THE FUTURE OF LAW ENFORCEMENT (2017); John Zerilli et al., *Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?*, AOP PHILOS. TECHNOL. 1 (2018); Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147–1224 (2016). Jane R. Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME LAW REVIEW 1 (2018) provide one of the few serious attempts to analyze the gaming issue, though they do not delve deeply into the question of when disclosure can and will lead to undesirable gaming.

[4] Trade secrecy claims, while ostensibly aimed at potential competitors, can also be used strategically to avoid accountability. Indeed, strategic assertions of secrecy to avoid accountability may begin with specialized vendors, who often cloak their algorithmic tools with secrecy, potentially avoiding accountability not only to decision subjects, but also to the ultimate decision-makers who rely on them. We mostly ignore these complications here to focus on strategic interactions between algorithm users/designers and decision subjects. We note as an aside, however, that claims that trade secrecy is needed to deter free riding competitors and incentivize innovation may be dubious or even pretextual when network effects or other first mover advantages are significant. See, e.g. Yafit Lev-Aretz and Katherine J. Strandburg, *Regulation and Innovation: Approaching Market Failure from Both Sides* (draft); Eli Siems, Nicholas Vincent and Katherine J. Strandburg and, *Trade Secrets and Markets for Evidential Technology* (draft); Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STANFORD

be misaligned with social welfare for at least three sorts of reasons: First, decision-makers may not adequately account for the intrinsic value of disclosure to decision subjects and society at large.[5] Second, decision-makers may fail to account for the social costs and benefits of particular sorts of inaccuracy and bias. Third, decision-makers may have self-serving, and strategic, incentives to hide the details of their decision-making bases and procedures from those to whom they are accountable, such as supervisors, government officials or the public at large. While opacity can sometimes prevent decision subjects from gaming the system, it can also mask socially undesirable algorithm design. Similarly, when decision-makers argue that opacity is necessary to avoid undesirable gaming they may be sincere – or they may be making such claims strategically. As a result, arguments for opacity based on a threat of gaming by decision subjects must be taken with a grain of salt, especially when made by government officials or by decision-makers subject to regulations that may find them liable, such as anti-discrimination laws or consumer protection regulations.

This Article focuses on a threshold analysis of the conditions under which the threat of "gaming" an algorithmic decision tool plausibly justifies non-disclosure of information about the algorithm. At bottom, a claim that disclosure will unleash gaming fails when the potential for socially undesirable gaming is low, regardless of whether it is made sincerely or strategically. For example, a law school's disclosure that it will only interview law school candidates with an LSAT above a certain threshold (a very simple "algorithm" using LSAT score as a proxy for likelihood of law school success) is likely to incentivize strategic behavior aimed at improving LSAT scores. This strategic behavior may or may not be socially undesirable, depending on whether the activities employed to improve LSAT scores actually improve a candidate's prospects of law school success.

While our inquiry is motivated by the recognition that decision-makers' incentives are often imperfectly aligned with social value, this threshold analysis allows us to make headway without delving deeply into decision-maker incentives. Our analysis suggests that, from a social perspective, the threat from "gaming" is seriously overstated by its routine invocation as an argument against disclosure. The resulting over-secrecy deprives society not only of disclosure's benefits to decision subjects, but also of the improvements in decision quality that could result when disclosure leads to better accountability.

LAW REVIEW 1343 (2018); Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA LAW REVIEW 54-None (2019).

[5] See, e.g. Barocas and Selbst; Strandburg

Insights from game theory, a method that studies strategic interactions among human decision-makers, can therefore be helpful in illuminating the algorithmic transparency question. Signaling theory, a branch of game theory which analyzes situations involving incomplete or asymmetric information (for example, an employer interviewing a job candidate), is particularly helpful.[6]

In earlier work, we identified general conditions that determine when strategic responses by decision subjects are likely.[7] Here, we review those earlier conclusions and then analyze some of the conditions that determine the extent to which such responses are likely to be socially desirable (or undesirable). Gaming is costly for society for two distinct reasons. First, engaging in gaming is costly for decision subjects in terms of time and effort that is wasted, rather than employed productively. Second, and more obviously, gaming is costly when it makes the proxy less informative, meaning that decision-makers must engage in further screening efforts. Decision subjects cannot meaningfully be said to "game the system," however, if they respond to disclosure by investing effort in changing their behavior in socially desirable ways. We address this question by framing gaming in terms of its likely effects on different sorts of mistaken decisions, given various relationships between the characteristics that would ideally inform decision-making bases and the algorithmic outputs that decision-makers use as proxies for them.

II. The Ubiquitous Gaming Trope

A. Gaming and Automated Decision-Making Algorithms

The issue of interest to us here arises from a common scenario: Decision subjects (or potential decision subjects) demand information about the bases for decisions that disadvantage them. Decision-makers respond that disclosure is "undesirable, such as when it discloses private information or permits tax cheats or terrorists to game the systems determining audits or security screening."[8] The force

---

[6] Michael Spence, *Job Market Signaling*, 87 THE QUARTERLY JOURNAL OF ECONOMICS 355 (1973). For applications to law, see, e.g., DOUGLAS G. BAIRD, ROBERT GERTNER & RANDAL PICKER, GAME THEORY AND THE LAW (Revised edition ed. 1998) chapter 4; NOLAN MCCARTY & ADAM MEIROWITZ, POLITICAL GAME THEORY (2014) chapter 8.

[7] Ignacio Cofone and Katherine Strandburg, Strategic Games and Algorithmic Secrecy, McGill L. J. (forthcoming 2019).

[8] Joshua A. Kroll et al., Accountable Algorithms, 165 U. Pa. L. Rev. 633, 638 (2017), p.633-34 (also stating at p. 639 that "The process for deciding which tax returns to audit, or whom to pull aside for secondary security screening at the airport, may need to be partly opaque to prevent tax cheats or terrorists from gaming the system. When the decision being regulated is a commercial one, such as an offer of credit, transparency may be undesirable because it defeats the legitimate

of the gaming argument depends, by implication, on an assertion that the social costs of gaming outweigh the benefits of disclosure. The specter of gaming is raised in a range of situations, yet the implied cost-benefit analysis is rarely spelled out in any detail.

For the most part, the "gaming the system" trope is employed in much the same way to justify keeping decision-making criteria secret whether those criteria are profiles implemented by human decision-makers or automated decision-making algorithms. Increasing reliance on complex models created by applying machine learning to "big data" has, however, introduced an additional wrinkle to the debate. Profiles applied by human decision-makers are always potentially understandable to decision subjects; the question is merely whether or not it is a good idea to divulge them. Machine-learning-based models, in contrast, may not be fully understandable even to decision-makers who employ them. Moreover, there are mathematically provable trade-offs between "accuracy" and "explainability" as they are commonly defined in the data science literature. The threat of gaming can be combined with this accuracy/explainability trade-off to argue for the acceptability of decision-making criteria that are not only kept secret from decision subjects and the public at large, but also unknown to decision-makers themselves. Conflating the issue of gaming with the question of algorithmic interpretability is a mistake, however. To begin with, the mathematical trade-off relates to a specialized definition of "accuracy" that is not equivalent to the decision-making performance that ultimately matters to society. Similarly, the trade-off involves a type and degree of "explainability" that may not be necessary to make disclosure effective in reining in self-serving decision-maker behavior. As Barocas and Selbst point out in a recent article addressing the issue of explanation, useful disclosure of information about a machine learning algorithm could take many forms.

In the discussion that follows, we do not attempt to tease apart the varieties of possible disclosure and their implications in any detail. We emphasize, however, that society's potential losses from decision subject gaming and society's potential benefits from better oversight of decision-makers both must be evaluated in light of the specific disclosures to be made.

B. Proxies, Disclosure and Gaming

---

protection of consumer data, commercial proprietary information, or trade secrets. Finally, when an explanation of how a rule operates requires disclosing the data under analysis and those data are private or sensitive (e.g., in adjudicating a commercial offer of credit, a lender reviews detailed financial information about the applicant), disclosure of the data may be undesirable or even legally barred.")

In prior work, we showed that the overall performance of a decision-making algorithm will be determined by both the noisiness of the proxies employed and the extent to which the algorithm is gamed.[9] These factors are not independent. As we have already discussed, strongly correlated proxies are generally difficult to game because there is usually some underlying reason for the strength of the correlation. Even a loose decision-making proxy can sometimes be relatively impervious to gaming, however.

First, disclosing an algorithm's use of unalterable features creates no potential for gaming, except in the rare situation when the decision subject is able to fake the input data without modifying the feature. Second, even when a feature is alterable in principle, there will be no gaming unless decision subjects decide that investing in altering that feature is cost-effective in light of the benefits of an improved decision outcome. Third, some features are of such obvious relevance to particular decisions that decision subjects will not need to be told that decision-makers are likely to take them into account. For such obviously relevant features, only a very detailed disclosure about how the feature plays into the decision can lead to increased gaming. Fourth, even if disclosure motivates a decision subject to invest in strategically modifying a particular feature, the result may not be socially undesirable gaming if modifying the feature improves the subject's true eligibility for a positive outcome. Developing good financial habits in hopes of getting a loan, studying hard to get a good GPA so as to get a good job and similar durable alterations leave the decision subject more deserving of a positive outcome than she was before. And finally, even in contexts where disclosure facilitates some gaming, it may also create opportunities for socially beneficial error correction by those for whom the proxy is a bad predictor of the characteristic of interest. This last point is particularly important when a proxy is systematically less accurate for some social sub-groups than for others.

C. When decision subjects can game the system

In summary, we have already argued that disclosure cannot seriously increase the threat of socially undesirable gaming unless several prerequisites are met:[10]

i) Decision-making proxies do not correlate well with the ideal decision-making criteria, so that there is enough "wiggle room" for gaming

---

[9] [cite]

[10] Cofone and Strandburg, supra note X

ii) The proposed disclosure must pertain to features (or feature data) that are (sufficiently) modifiable by decision subjects
iii) Modifying those features must be cost-effective
iv) Modifying those features must improve the proxy without improving the decision subject's true eligibility for a beneficial decision
v) If a proposed disclosure requirement does not meet these prerequisites, gaming arguments should be discounted.

These prerequisites can be used to create a framework for analyzing when policymakers should mandate disclosure of decision-making algorithms and proxies. Beyond these basic prerequisites, the policy analysis should also account not only for the potential costs of gaming, but also for the potential benefits of disclosure, including the possibility of decision subject "error correction," where strategically altering the proxy compensates for noise or bias in its correspondence to the ideal decision-making criteria. To flesh out the possibilities, the rest of the paper considers the costs and benefits of disclosure of decision subjects' strategic responses to disclosure are likely to vary depending on the sorts of errors that a proxy tends to make.

III. The Principal-agent Problem

A. Decision-Makers as Strategic Actors

Disclosure of the proxies and procedures used in decision-making often has the potential to confer significant social benefits by promoting accountability, improving decision accuracy, deterring or exposing bias, arbitrariness and unfairness, permitting decision subjects to challenge the factual or other bases for erroneous decisions and to undertake the socially beneficial strategic behaviors discussed in the previous Part. Decision-makers also need not sit back and let gaming proceed unchallenged. They can respond to the threat of gaming by adopting proxies that are less easy to game and can sometimes use selective disclosure to discourage subjects from attempting to game the system.

If decision-makers could be trusted to have society's interests at heart, they presumably would weigh the social benefits of disclosure against the potential costs of socially undesirable gaming and make socially optimal decisions about whether, and in what detail, decision-making algorithms should be disclosed. But decision subjects are not the only ones who can play games. Many of the social benefits of disclosure arise precisely because disclosure addresses conflicts between decision-maker incentives and the public good. The threat of gaming can itself be wielded

strategically by decision-makers seeking to avoid accountability, cut corners, cover up bias or otherwise place their own interests above those of society at large.

B. Decision-Makers as Imperfect Agents of Society's Interests

The fact that decision-makers are imperfect agents of society's interests is hardly news. There is a large literature associated with the problem of "public choice": the ways in which the private interests of government actors can distort their behavior away from the public interest they have been appointed to serve. Explanation of government decision-making is a core requirement of procedural due process that is intended, at least in part, as an accountability mechanism. Consequential private sector decisions are also subject to legal disclosure requirements. For example, fair credit laws demand a certain level of disclosure to applicants about the bases for loan denials. In other arenas, such as employment and housing, while the law does not require disclosure of decision-making criteria, it does prohibit reliance on certain characteristics, such as race, gender, age and disability.

The principal-agent problem in government decision-making manifests in various ways. Government decision-makers may shirk, for example, investing less in decision-making than would be socially optimal. Or they may over-emphasize certain kinds of mistakes and under-emphasize others, as when an elected judge over-weights the reputational risk associated with releasing defendants compared to the social and individual costs of unnecessary detention. Private decision-makers serve less obviously as agents of the public interest, but in contexts such as employment, education, housing and credit, their decisions about issues such as how much care to take to avoid bias and discrimination may also have significant externalities affecting the public interest. Automating some or all of the decision-making process is not a silver bullet for avoiding such principal-agent problems; it simply moves them upstream to the point at which the automated process is designed or procured.

Tension between society's interests and decision-makers' personal interests not only affects the way decisions are made, but also gives decision-makers incentives to avoid accountability and embrace opacity. When decision-makers value the private benefits afforded by opacity, they can be expected to exaggerate the threat that disclosure will degrade decision-making performance by allowing decision subjects to game the system. This is not to suggest that decision-makers are unconcerned with making sound decisions or that their warnings about the potential for gaming should go unheeded. The point is only that, when push comes to shove, decision-makers may not make socially optimal trade-offs between

investments in accuracy and the social costs of various sorts of errors and may exaggerate the threat of gaming in order to protect their private interests.

C. Strengthening the Proxy to Avoid Gaming

As noted above, weak proxies are more likely than strong proxies to be gameable. As a result, weak proxies and gaming will often go hand in hand. Decision-makers (or, perhaps more to the point, algorithm providers) can respond to the association between weak proxies and gaming by attempting to hide the fact that their decision procedures are not based on robust proxies. The threat of gaming, which will often be real for algorithms employing weak proxies, provides a convenient excuse for such opacity.

But opacity is usually not the only available way for decision-makers to discourage gaming; instead, decision-makers can often opt to devise stronger proxies, thus simultaneously improving decision performance and making gaming more difficult. By adopting -- and then disclosing -- better proxies, decision-makers can sometimes encourage decision subjects to invest in developing durable features that improve their qualifications for positive decision outcomes, often simultaneously producing better results for decision-makers. Suppose a software company had been screening potential employees by looking at data about subscriptions to the top PC and Mac magazines and websites. This proxy is likely to be easily gameable, in part because it is a weak proxy for software engineering skills. If the company starts basing its screening on college grades in software engineering classes instead – and if college grades are a reasonably sound proxy for performance -- the company can benefit from both adopting and disclosing its new criterion. Gaming a strong proxy tends to be costly for decision subjects. The only way to game the good grades proxy without putting in the hard work of studying and learning the material is to fake the grades. But grades are easily verifiable by requesting transcripts from the college or university. Hacking into the university system to fake one's transcript is likely to be a risky and costly strategy (even for software engineers), and thus rarely cost-effective. If anything, disclosing the good grades criterion is likely to benefit the employer by incentivizing more potential employees to invest in obtaining good grades and, presumably, learning software engineering skills.

To summarize, the threat of gaming, coupled with a disclosure requirement can motivate decision-makers to devise and adopt better proxies for the ideal decision-making criteria. Unless upgrading the proxy is too costly, this strategic response can be beneficial for decision subjects, decision-makers and society overall.

D. Credible Threats and Strategic Disclosure

Under some circumstances, decision-makers can use selective disclosure strategically to attempt to convince decision subjects that gaming is more costly than it is. To see this, imagine someone who makes a practice of robbing ATMs. I If security cameras are visible, the thief's best strategy to gaming the surveillance system is to avoid them. Authorities might attempt to combat this sort of gaming by hiding the cameras, but announcing that they have been installed at all ATMs. As long as the cameras are hidden, however, another selective disclosure strategy may be possible. Rather than installing cameras at all ATMs, authorities can install them in a few locations and use vague pronouncements or even outright misrepresentations to give the impression that cameras in ATMs are more common than they really are. If effective, such a combination of disclosure and opacity rules out the gaming strategy, while convincing ATM thieves that the proxy is more robust than it is, thereby incentivizing compliance.

This interaction between decision-makers and decision subjects can be stylized in terms of the economic concept of credible threats. A decision-maker may declare the cost of gaming a system in the hopes that as few people as possible will try to game it. The declaration "this is how difficult our proxies are to fake" is essentially a threat that anyone trying to game the system will be unsuccessful. This sort of threat strategy will deter gaming only if the threat is *credible*.

After a vague announcement threatening that ATM thieves will be caught by hidden cameras, people might not respond by giving up the crime. Instead, they (at least collectively) might test the credibility of the threat by reducing activity levels and seeing whether they are caught. If thieves rarely or never get caught at particular ATMs, the vague disclosure about cameras may no longer serve as a credible threat. By disclosing details about its proxy, (i.e. how many cameras there are and where), a decision-maker allows decision subjects to verify for themselves how costly it would be to game the system, making the threat implicit in the disclosure highly credible.

Because detailed disclosure exposes the true costs of gaming, it may or may not be a sufficient deterrent. Decision-makers may thus be tempted to try to exaggerate the threat. If the proxy is not disclosed in sufficient detail, however, decision subjects may suspect that the threat is exaggerated (e.g. if law enforcement announces that there are lots of hidden cameras, but people see no evidence of their existence) and the threat may no longer be credible. As a result, while the strategy of exaggerating the costs of gaming may initially seem appealing to decision-makers, disclosing enough information about the proxy to allow subjects to verify that the threat is credible may often be a more effective decision-maker strategy.

Such a "revealing the proxy" strategy is beneficial whenever the proxy is costly to fake and is strongly correlated with the underlying decision bases. When the proxy is costly to fake, decision subjects will be incentivized to alter their behavior to improve their decision outcomes, rather than attempting to game the system. When the proxy is strongly correlated, those alterations are likely to improve the subjects' underlying eligibility for a positive outcome. In the ATM camera example, if cameras are everywhere, people will be more likely to comply. If faking is cheap, however, (e.g. there are few cameras or they can be easily fooled by wearing a hoodie or sunglasses), decision-makers may choose to hide the proxy to avoid gaming, returning to the situation analyzed in the previous section.[11]

IV. Law, Compliance, Discretion and Gaming

While much of the analysis of the argument for opacity based on gaming applies to both government and private sector decision-making, government decision-making raises some distinctive issues, which are discussed in this Part.

A. Accuracy and Legitimacy in Legal Proxies

Legal rules are sometimes framed explicitly in terms of indirect proxies for policy targets. For example, licensing exams, such as bar exams, are used as proxies for professional competence.[12] When legal rules employing proxies have been adopted through legitimate democratic processes, and assuming they meet constitutional requirements, they have procedural legitimacy that proxies adopted by other means do not. Even when legal rules are only proxies for the underlying aims of legislators or administrative agencies, governmental decision-makers may, indeed must, base their decisions on the enacted terms of those proxies when assessing legally determinative elements or factors. Relying on legally enacted proxies, even if they are indirect, is thus an entirely different matter from relying on indirect, secret proxies.[13] In other words, while a judge can decide that circumstantial evidence in a case meets the relevant standard of proof of the elements of a tort or crime, that judge may not replace that determination with a secret checklist of indirect proxies for those elements.[14]

Moreover, legal requirements must be publicly disclosed and citizens are obligated only to comply with the law as implemented through those proxies; they

---

[11] See SectionX.

[12] [cite]

[13] [cite]

[14] Indeed, an explanation requirement might be one tool for detecting when that has happened.

are not required to independently act in accordance with the policy goals of the law. This complies with the generally accepted standard that government decisions must be based on legally determined and publicly disclosed factors.[15] This limitation on the obligation to comply applies even when citizens knowingly and intentionally "game" the legal standard to subvert its intent. For example, in tax law, this is the difference between elusion (complying with tax rules while subverting legislative intent) and evasion (breaking tax rules to subvert legislative intent). It is therefore inconsistent to keep legal requirements secret for fear that citizens will game them. In fact, explanation of decision-making is the most basic requirement of due process. Concerns about citizens "gaming the system" in an *illegitimate* way, thus can only arise in arenas where government officials can legitimately exercise discretion in choosing to rely on proxies.

The limited scope of the law's legitimate concerns about gaming by citizens means that, when discussing the potential for gaming of government decision-making, there is a distinction between baseline legal rules and algorithms for governing enforcement and investigation.

B. Baseline Rules and Enforcement Rules

Take the example of drug enforcement. Drug enforcement officials use a variety of (arguably dubious) "profiles" to determine whom to target for investigation.[16] In the use of such profiles, the enforcement method is secret, but the substantive rule is not. For example, the substantive rule against drug trafficking is publicly disclosed. At an eventual trial, the evidentiary proxies for a finding of drug trafficking must also be disclosed to the defendant and to the public: for example, witness testimony, fingerprints and DNA from the suspect on drug packages, and a geolocation pattern that follows the drug's path.[17] The decision to investigate and enforce the rule against a defendant are, however, made on a basis that is not revealed to the defendant in advance and may not have to be disclosed to the defendant or the public unless there are constitutional concerns.

Similarly, the Internal Revenue Service (IRS) uses a combination of random sampling and proxies to direct its targeting of tax audits.[18] While tax professionals attempt to guess these proxies to assist their clients, the IRS attempts to keep its

---

[15] [cite]

[16] [cite]

[17] There is, however, controversy over the extent to which the underpinnings of these evidentiary proxies must be disclosed. See, e.g., Siems, Vincent and Strandburg.

[18] [cite]

proxies secret,[19] ostensibly so that taxpayers cannot "game the system" by submitting fraudulent tax returns. Thus, while the tax laws and IRS rules are publicly disclosed so that, at least in theory, people may inform themselves about what they are obligated to declare, the enforcement rules are not. The IRS tells citizens what to do, but it does not tell them how it finds out whether they are doing it. In other words, it is transparent about its baseline rules but not about its enforcement rules.

In these situations, the government seeks to use relatively easily detectable indirect proxies as tools for directing investigations into illegal behavior that is difficult to detect. This difference between baseline rules and enforcement traces what Dan Cohen calls an acoustic separation between "conduct rules" and what he calls "decision rules."[20] Conduct rules speak to decision-subjects telling them how to behave, while decision rules, in his terminology, speak to officials who enforce conduct rules (decision-makers), telling them how to react when a subject breaks a conduct rule.[21] For government officials, the duties of disclosure to decision subjects are higher for conduct rules than they are for decision rules. Law's authority, some would argue, justifies concealing decision rules from subjects.[22] The corollary of this justification is that such concealment is only justified when backed by law's authority; this is, when the decision-makers are public officials with procedural legitimacy.

Other types of government decisions, such as, sentencing and pre-trial detention, inevitably leave space for decision-maker discretion because they depend, at least in part, on an assessment of "character," "propensity," or other general or probabilistic factors that cannot be tied tightly to specific factual elements.[23] In such decisions, algorithmic proxies (whether automated or simply articulated to decision-makers) are often proposed as a means of directing and cabining decision-maker discretion.[24] Depending on the proxies that are used, it might be possible for subjects of such decisions to "game the system" by altering

---

[19] [add source]

[20] Meir Dan-Cohen, *Decision Rules and Conduct Rules: On Acoustic Separation in Criminal Law*, 97 HARVARD LAW REVIEW 625–677 (1984).

[21] *Id.*

[22] Raz, "Authority, Law, and Morality" in Raz, *Ethics in the Public Domain: Essays in the Morality of Law and Politics* (Oxford: Oxford University Press, 1994).

[23] [cite]

[24] [cite]

their surface behavior, without changing the underlying feature that the decision-maker seeks to assess.[25]

C. The Choice between Secret and Disclosed Proxies for Investigative and other Discretionary Decisions

The use of secret proxies for making investigative decisions is defended in part by concerns about gaming and in part by the ostensibly low cost of being investigated by mistake.[26] Arguably, government officials often cannot avoid using inexpensive, observable proxies to target investigative effort. When that is the case, it may be preferable to cabin discretion and avoid implicit bias by using explicit proxies, rather than the "black box" of human judgement and intuition. Moreover, the use of explicit proxies, such as machine-learning-based algorithms, allows for the incorporation of empirical evidence about what kinds of proxies are most accurate.

This sort of argument for using explicit proxies is, in essence, an argument *in favor* of greater disclosure of the bases for investigative and other discretionary decisions; the objection is not to disclosure, per se, but to disclose *to the potential subjects of investigations*. The gaming argument applies only to *public* disclosure of how investigations are targeted. Moreover, our analysis here and in our previous article demonstrates that any assumption that disclosing the proxies used in investigative and other discretionary contexts will lead to socially undesirable gaming is far too facile. Gaming risks vary and must be weighed in light of the importance of accountability and avoiding bias in these government activities.

V. Normative Considerations: When Should Disclosure Be Required?

A. Decision Subject Strategies and Types of Decision-Making Errors

Our discussion has so far treated all decision errors as equivalent. This approach is consistent with the way that the term "accuracy" is used in data science to indicate the total error rate. In many decision contexts, however, there are important, policy-relevant differences in the social costs of different sorts of errors and in the trade-offs between disclosure's social value and the threat of gaming it creates. To explore these variations, this Part decision outcomes in terms of three binaries: "action" or "inaction," "beneficial" or "detrimental" and "correct" or

---

[25] As discussed above, however, decision-makers can also "game this system." [cross-reference]

[26] [cite]

"mistaken." [27] In describing a decision as an "action" or "inaction," we adopt the decision-maker's perspective. For example, in the hiring context, a decision to hire is an "action," while a decision not to hire is an "inaction." In the law enforcement context, a decision to arrest or investigate an individual is an "action," while a decision not to arrest or investigate is an "inaction." A decision outcome is either "beneficial" or "detrimental" from the decision subject's perspective. In the employment example, an "action" decision is also "beneficial" while, in the law enforcement example, an "action" decision is "detrimental." The terms "correct" and "mistaken" are used to connote the relationship between the outcome of the decision-making process and the ideal decision.

Thus, in the employment context, a "correct action" decision results in the hiring of a qualified candidate; a "mistaken action" decision results in the hiring of an unqualified candidate; a "correct inaction" decision means that an unqualified candidate was not hired; and a "mistaken inaction" decision means that a qualified candidate was not hired. A similar analysis applies in the law enforcement context, except that an "action" outcome is detrimental to the decision subject.

Assuming the basic prerequisites of feasibility and cost-effectiveness are met, decision subjects anticipating detrimental decisions (whether correct or mistaken) will respond strategically to disclosure by modifying their features (or feature data) to produce beneficial decisions. Decision subjects anticipating beneficial decisions (whether correct or mistaken) will sit tight. Socially undesirable gaming occurs when an individual who would otherwise have received a correct detrimental decision alters her features to so as to receive a mistaken beneficial decision.

However, disclosure might also facilitate two sorts of socially desirable responses: i) by altering her features an individual who would otherwise have received a correct detrimental decision might end up complying with the ideal decision criteria, thus qualifying herself for a correct beneficial decision; or ii) an individual anticipating a mistaken detrimental decision might be able to alter her features to correct the error in the proxy and obtain a correct beneficial decision. The possibilities are illustrated by the table below. Any accounting of the social

[27] While not all decision outcomes can be described as "action" or "failure to act," many important sorts of decision are of this sort and, in any event, framing the discussion in terms of binary outcomes is sufficient to provide insights into the issues involved. Also, we forego the usual "false positive" and "false negative" terminology because it becomes confusing when we need to refer to both decisionmaker and decision subject perspectives.

costs and benefits of disclosure should offset the costs of any socially undesirable gaming with the benefits of the latter two responses.

Pre\Post Disclosure Decision	Correct beneficial	Mistaken beneficial
Correct detrimental	Compliance (Socially Desirable)	Gaming (Socially Undesirable)
Mistaken detrimental	Error Correction (Socially Desirable)	N/A

The next few sections analyze the likelihood that a convincing case can be made that opacity is necessary to thwart gaming for proxies with various error profiles. Social tolerance for different sorts of errors also varies with context. In the law enforcement context, for example, mistaken convictions (actions) are afforded more weight than mistaken acquittals (inactions).

B. Highly Accurate Proxies: Low Numbers of both Mistaken Actions and Mistaken Inactions

Consider an algorithmic decision-making tool that produces a set of decisions for which rates of both mistaken actions and mistaken inactions (and thus, necessarily, mistaken beneficial and mistaken detrimental rates) are low. In other words, the algorithm uses a highly accurate proxy that is displayed by nearly everyone that a decision-maker with access to the ideal criteria would select (low mistaken inactions) and almost no one that a decision-maker would not ideally select (low mistaken actions). For example, suppose a CV screening algorithm makes very few mistakes when using involvement in charitable activities as a proxy for identifying police officer or teacher candidates who are dedicated to community service. Because such a highly accurate proxy is socially valuable, there would be a good argument for keeping such an algorithm opaque if disclosure would facilitate significant gaming, which would correspond, in our example, to getting involved in charitable activities only as a means to land a job interview.

In most cases, however, the costs of gaming such an algorithm are likely to be high because this error profile usually means that the proxy correlates strongly with the ideal decision criteria, perhaps because they are causally related to one another or to some third factor, because they are synergistic in some way or because there is some contextual or social reason that those who meet some ideal criterion tend to develop a preference for the proxy behavior. Modifying such a proxy without changing the ideal outcome, as would be required to game the system, will ordinarily be difficult and expensive. For example, engaging in many charitable activities is likely to be more enjoyable and valuable to those with a propensity for community service, an example of a synergistic effect. Moreover, it seems reasonably likely that engaging in charitable activities encourages a propensity for community service.

An exception to the general rule that highly accurate proxies are ungameable can arise when a proxy correlates with the ideal criteria but in a temporary or incidental way. Suppose, for example, that law enforcement authorities have observed that all members of a drug gang, and almost nobody else, meet frequently at a house. Going to that house might then be a quite accurate proxy, with low false negatives and low false positives, for involvement in that drug gang, and thus in drug trafficking. A decision to target that house for surveillance would be both well-justified and likely to bear fruit. If, however, the gang learns that the police are wise to this meeting place, the gang's size and social structure might allow it to switch meeting places relatively easily. This example illustrates the relatively unique type of situation in which a proxy is highly accurate, but relatively cheaply gameable.

Not coincidentally, the gameable but high accurate proxy has characteristics akin to the particularity that is required by the Fourth Amendment and most surveillance statutes. This type of particularized proxy is common in law enforcement. In fact, the warrant process is itself a proxy-based decision-making system. Notably, to obtain a warrant, law enforcement officials may keep the proxy secret from the target but must explain the non-arbitrary and particularized connection between the proxy and the suspected crime to a judicial magistrate. We speculate that particularity is a necessary (though not sufficient) characteristic of highly accurate, but gameable, proxies. Outside of these special cases, highly accurate algorithms can ordinarily be disclosed without incurring significant gaming costs.

In sum, highly accurate proxies provide good reasons for opacity in the relatively rare situations when they are gameable because the correlation between the proxy and feature is essentially incidental or temporary. The Fourth

Amendment's particularity requirement may aim at protecting this specific sort of proxy-characteristic relationship.

B. Noisy Proxies: High Numbers of both Mistaken Actions and Mistaken Inactions

At the other end of the spectrum, consider a proxy for which rates of both mistaken action and mistaken inaction (and, necessarily, mistaken beneficial and mistaken detrimental rates) are high, meaning that relying on the proxy results in selecting large numbers of the wrong individuals, while also missing lots of individuals who should have been selected. For example, imagine using a particular religious identity as a proxy to screen for terrorists at airports. Religion and participation in terrorist activities do not correlate well, so such proxy would produce both high mistaken actions (in this case, mistaken detrimentals) for non-terrorist members of that religion and high mistaken actions (in this case, mistaken beneficials) for terrorist non-members of that religion

Using such an inaccurate proxy would contribute marginally, at best, to the ultimate goal of making the right decision. Nonetheless, using such a proxy might be socially justified if the costs of employing a more accurate proxy would outweigh the total error costs incurred by decision-makers and decision subjects and if there are no significant distributional concerns. That is unlikely for terrorism, but could be the case for a less consequential decision. Any example of such a proxy is contestable, but the use of LSAT scores as an initial screener for law school admissions might be one.[28] Of course, applicants who receive "mistaken detrimental" decisions, in that they are screened out of their preferred schools although they would have been successful students there, would not characterize the error cost as small. Nonetheless, the overall social costs of inaccurate screening might be lower than the costs to law schools of investing in significantly more accurate proxies, at least at the initial cut-off stage.

Even if a weak proxy would be gameable, opacity thus may not be socially optimal. Highly inaccurate proxies are of relatively low social value to begin with, so that gaming may not significantly undermine their value. Since the mistaken detrimental rate is high, disclosure may facilitate not only socially undesirable gaming, but also significant error correction. Since the system already tolerates a

[28] This example is contestable for various reasons, including the possibility that LSAT score produces high false negatives, but low false positives. But bear with us for purposes of illustration.

high level of mistaken beneficial decisions, gaming may not make things too much worse.

Keeping a highly inaccurate proxy secret to stave off gaming deprives decision subjects of the opportunity to identify and complain about the inaccuracy generally and, more importantly, about any disparate impacts that it may cause. For example, and not uncommonly, the mistaken detrimental and mistaken beneficial decisions resulting from a low accuracy proxy may be distributed unevenly, so that the burdens of loose targeting are borne disproportionately by some sub-group. For example, high income individuals might be over-represented in the mistaken beneficial group because they can afford LSAT prep courses and specialized tutoring, while low income individuals might be over-represented in the mistaken detrimental group because they cannot take advantage of LSAT-specific training. Disclosure can thus reveal not only noise, but socially concerning disparity.

Moreover, while the use of inaccurate proxies is sometimes justifiable on cost-benefit grounds, it may also be a sign of shirking by decision-makers who do not bear the full costs of their errors. For that reason, arguments for keeping highly inaccurate proxies secret to avoid gaming should also be viewed skeptically, at least for decisions that have important consequences for decision subjects (which are our primary focus).

In sum, highly inaccurate proxies are generally of low social value and, therefore, even when they are gameable, the aggregate social cost of gaming will be low. Moreover, these are a suspect category in terms of agency problems and disparate impact. Therefore, transparency is optimal in important decision contexts. Even for less significant decisions, it will often be desirable to require some level of disclosure to provide accountability, while keep gaming in check.

VI. Normative Considerations: Asymmetric Proxies

A. Low Mistaken Actions, but High Mistaken Inactions

Proxies for which mistaken action and inaction rates (and hence mistaken beneficial and mistaken detrimental rates) are asymmetric are accurate enough to be valuable, but can impose significant costs on decision subjects. In this section, we analyze proxies that have low rates of mistaken action, but high rates of mistaken inaction. The following section considers the reverse.

At first glance, a proxy with a low false positive rate and a high false negative rate might seem unproblematic from a social perspective. Its low false positive rate means that the individuals selected, whether for a benefit such as employment or a burden such as law enforcement investigation, are generally

"deserving" of the outcome that they receive. But what of the high false negative rate?

There are two distinct situations to consider. In the first type, the nature of the decision, or resource) constraints, means that the benefits (or burdens) of "action" decisions will ultimately be obtained (or borne) by a limited number of individuals. Employers, for example, cannot hire unlimited numbers of workers; schools cannot admit unlimited numbers of students.[29] In situations of this sort, mistaken inactions are inevitable and, as long as the mistaken action rate is low, the costs associated with mistaken inactions are likely as low as can be expected (i.e. mistaken inactions cannot be eliminated by using more accurate proxies). Of course, even here agency problems could lead to a biased distribution of mistakes.

The situation is different when the number of actions outcomes is unlimited. In that case, each subject receives an independent determination of whether he or she receives benefits (or is subjected to burdens).[30] In such situations, a high rate of mistaken inactions can inflict significant social costs even when there are few mistaken actions: while the previous decision with a fixed number of spots was a zero-sum game, this is a positive-sum game.

When an action outcome is beneficial for the decision subject, mistaken inactions are mistaken detrimental decisions, with costs borne primarily by decision subjects. An example of this is credit scores: a falsely low credit score underestimating one's ability to borrow is detrimental, and a falsely high credit score overestimating one's ability to borrow is beneficial. On the other hand, when an action outcomes is detrimental for the decision subject, mistaken inactions are mistaken beneficial decisions. An example of this is parole decisions: a falsely low risk score underestimating one's propensity to recidivate is beneficial, and a falsely high risk score overestimating one's propensity to recidivate is detrimental. While decision subjects benefit from mistaken beneficial decisions, they may impose costs on decision-makers and society at large. If, for example, a mistaken inaction means releasing a dangerous criminal, society would pay a cost. Similarly, failing

[29] Most examples of this situation involve decisions where there is a limited number of "spots" to fill. In those cases, a positive decision is a benefit and a negative decision is a burden. However, the reverse is also possible. For example, law enforcement officials ordinarily must choose where to target limited resources: mistakenly being investigated is burdensome and mistakenly not being investigated is beneficial for decision subjects.

[30] Sometimes there are resource constraints over the long term, even when decisions are nominally independent. For example, individual decisions whether to award various sorts of government benefits or to send individuals to prison have cumulative resource implications that undoubtedly feed back into decision criteria. For the moment, we ignore these complications, though we note that they can exacerbate principal-agent issues.

to investigate a criminal or terrorist produces costs borne by society that may be large.

Moreover, when the number of actions outcomes is unlimited, yet the number of mistaken inactions is large, we cannot be assured that decision-makers are acting as faithful agents for the public interest: like a highly inaccurate proxy, this scenario creates a suspect category for agency problems. Even when all or most of the action decisions are correct, the decision-maker may simply be taking too little action. . Perhaps mistaken actions are more visible than mistaken inactions and the decision-maker is risk averse. Perhaps action outcomes are more expensive for the decision-maker to handle, despite sufficient social benefit. Or perhaps the decision-maker is simply shirking the task of devising a proxy that could correctly identify more of the individuals currently receiving mistaken inaction outcomes.

Here, too, it is possible that mistaken action and inaction outcomes are distributed in a biased fashion. If there is no bias, mistaken actions and inactions should be distributed essentially randomly among the population. Bias in the distribution can arise in at least two distinct ways. It might be that the proxy accurately picks out decision subjects deserving of "action" in one population group, but not another. In that case, the number of mistaken actions might be small, even though the proxy is biased. For example, law enforcement officials might have created an extremely accurate profile of urban African American drug dealers that is a terrible proxy for identifying white suburban drug dealers. Or perhaps our proxy is good at identifying city dwellers who acquire guns for protection in violation of gun laws, but bad at identifying rural NRA members who are violating the same laws. Perhaps it does a great job of identifying men who would make good prosecutors, but a bad job of identifying women who would be good at the job. This sort of distributional disparity is problematic even though the rate of mistaken action is low – and even in cases where the number of "slots" for positive decisions is limited.

A proxy might also have a disparate impact, not because the profiles differ between sub-groups, but because the proxy is "polluted" by irrelevant factors that separate individuals into "correct action" and "mistaken inaction" outcomes in a way that is correlated with suspect distinctions along lines of race, gender, and the like. Here again, the proxy is problematic even when the rate of mistaken action outcomes is low. Imagine, for example, that the IRS uses a five-factor checklist as a proxy for "likely tax evader," in which four of the factors are truly relevant to identifying tax evaders, while the fifth factor is "living in a low-income neighborhood." The five-factor proxy will have a low false positive rate, since most

individuals targeted for investigation will be tax evaders, but a high false negative rate, since it misses tax evaders who do not live in low-income neighborhoods.

Disclosing proxies with low rates of mistaken action, but high rates of mistaken inaction would help to identify such problems of bias, but will it lead to costly gaming? Most of the time, if a feature separates some identifiable sub-group of the population from others, that feature is likely to be costly for members of that sub-group to change, even if the feature is not immutable. Gaming costs, as well as error correction costs, associated with "faking" one's sub-group membership are generally high. We have already suggested that proxies should be disclosed when gaming costs are high, in part for the very purpose of detecting and eliminating biased proxies.[31]

However, there is still the possibility that disclosure might facilitate gaming by members of the favored group who would otherwise receive detrimental decisions. Despite this risk, in contexts where it is plausible that bias accounts for the asymmetry between false positive and false negative rates, the need to detect bias will generally outweigh the potential costs of gaming. It will usually be socially preferable to require at least some sorts of disclosure so that bias can be detected. If within-group gaming is a serious concern, it may be possible to design the disclosure to illuminate potential bias while keeping enough details secret to deter socially undesirable gaming. Alternatively, of course, decision-makers could be required to use more accurate proxies.

As in the case of highly accurate proxies, particularized proxies may be exceptions to a general preference for disclosure where there are high social costs of gaming. Recall the hypothetical above, where law enforcement guesses whether someone is a member of a drug gang depending on presence at a particular house. Viewed only as applied to the particular gang, that proxy had a low false negative rate, since we constructed the hypothetical so that all members of the gang frequented the meeting location. Viewed more broadly, however, such a proxy would have many false negatives, since drug traffickers in other gangs, other cities, and so on would not frequent that meeting place. If the gang can easily find another meeting place, the costs of "gaming" this proxy will be low. When a proxy is gameable because of its particularity, we should be less concerned that it produces false negatives. Even then, however, it may be socially desirable to provide disclosure to some independent source of accountability, as we do for warrants, to keep an eye out for bias.

---

[31] [cross reference]

In sum, whether proxies with low rates of mistaken action outcomes and rates of mistaken inaction should be disclosed depends on the balance between the aggregate cost of gaming and the social value of accountability. This balance will be decision-dependent.

B. Low Numbers of Mistaken Inactions and High Numbers of Mistaken Actions

Finally, there is the possibility that a proxy produces few mistaken actions, but many mistaken inactions. Imagine two examples for this scenario. If an "action" decision is burdensome, as it is in the law enforcement context when the police decides whether to investigate someone, a proxy with a high rate of mistaken actions imposes costs on innocent individuals, that are mistakenly investigated in order to ensure that all guilty individuals are correctly investigated (low mistaken inaction rate). If an "action" decision is beneficial, as it is in the employment context when an employer decides whether to interview someone, a proxy with a low rate of mistaken inaction, but a low rate of mistaken action benefits some undeserving individuals who are interviewed even though they are not fit for the job, in order to ensure that all deserving individuals receive interviews.

As in the scenario with high rates of mistaken action outcomes and low rates of mistaken inaction outcomes, there are two situations: when the number of slots is limited and when it is unlimited. When the numbers of positive and negative slots are unlimited, this scenario is the flip side of the equivalent scenario discussed in the previous section and the analysis is applicable: the scenario is suspect for agency problems because the decision-maker may not fully absorb the social costs of mistaken decisions. Imagine, for example, an algorithm that decides whether a child will be taken out of his or her home and placed in foster care. Socially, we want families to have the benefit of the doubt and children to be taken out of their homes only if there is relative certainty that they may face violence or unsafety by staying there (a standard somewhere in between more likely than not and beyond reasonable doubt). However, children mistakenly taken out of their homes (mistaken actions) are relatively invisible to the broader public, and even to decision-makers, since there is no counterfactual available. Thus, decision-makers are relatively insensitive to the social costs of mistaken actions in this context. Every child mistakenly kept in their home (mistaken inaction), however, is a child that will be a victim of violence that draws media attention and is visible to and costly to the decision-maker. Thus, decision-makers have incentives to tolerate proxies with high rates of mistaken action as long as they have low rates of mistaken inaction, imposing externalities on children and families who are unnecessarily separated and on society at large.

On the other hand, the symmetry with the previous sections analysis does not trace as well when the number of "action" slots (whether beneficial or detrimental from the decision subject's perspective) is limited. The difference arises because, while a high rate of mistaken inactions is often of little concern to decision-makers when the number of "action" slots is limited, decision-makers are burdened by mistaken actions in that situation and will likely engage in costly additional screening to eliminate them. Employers, for example, are unlikely to be satisfied with a proxy that lets a high number of unqualified individuals through, even if it rarely screens out a qualified candidate because they will have to do something to further whittle down the candidate pool. Auditing individuals that a proxy identifies as likely tax evaders or surveilling individuals that a proxy deems likely to commit crimes similarly impose costs on decision makers. The additional screening may also impose costs on both correctly and mistakenly selected decision subjects, but as long as selection is potentially beneficial for decision subjects they may find the costs of additional screening well worth incurring. In any event, for beneficial decisions, subjects often have a choice about whether to incur the costs of additional screening. In law enforcement and other investigative and enforcement contexts, however, those mistakenly selected by a proxy are innocent individuals who have no choice but to bear the burdens associated with being investigated or even punished.

When the individual burden of being incorrectly targeted (mistaken action outcome) is high compared to the social cost of incomplete enforcement (mistaken inaction), this type of proxy may be socially unacceptable, even if decision-makers are willing to be the costs of additional screening. For example, because the burden of false imprisonment is extremely high, criminal law takes the view that it is better to have ten guilty people go free rather than punish one innocent person. In civil and administrative contexts, society tolerates much higher rates of mistaken actions, presumably because it considers that the social cost of mistaken civil liability is lower than that of criminal punishment and worth tolerating to obtain a lower rate of mistaken finding of non-liability.

Different standards of proof respond to concerns about the expected costs of mistaken action and inaction outcomes.[32] The standard in criminal law is beyond reasonable doubt, which aims to lower the number of mistaken action outcome by tolerating a relatively higher number mistaken inaction outcomes. The standard in civil suits is preponderance of the evidence, with higher rates of mistaken actions and lower rates of mistaken actions, presumably because the private and social costs

[32] [add cite]

of falsely imprisoning someone are much higher than those of mistakenly allocating damages. For example, if a proxy indicated a probability between 50% and 90% that an individual had engaged in sexual violence, legal standards of proof reflect the judgement that it would be unreasonable to imprison them, but reasonable to require them to pay damages to the victim.[33] In an enforcement or investigative context, we often tolerate a high initial false positive rate when the burden of being investigated is deemed relatively low. In the tax context, for example, it is considered acceptable to audit rather large numbers of taxpayers who turn out to be innocent of tax evasion. The standard for a wiretap, on the other hand, is considerably more stringent.

However, one should be concerned about proxies that mistakenly impose investigative burdens on large numbers of individuals when those individuals are predominately members of some sub-group of the population. In such a situation, the disparate impact may outweigh the benefits of better enforcement and the political process may be unlikely to sufficiently account for those burdens. Disclosure of such a proxy may be necessary to discover such a disparate impact. Recall, for example, the example of the decision-maker whose goal is to investigate terrorism. Imagine she assumes that the majority of terrorism is perpetrated by Muslims and therefore uses being Muslim as a proxy for being a terrorist, therefore imposing the burdens of investigation on large numbers of innocent Muslims. Even if were true (as often incorrectly assumed), that the majority of terrorism is perpetrated by Muslims, such a proxy would produce an extremely high rate of mistaken actions (investigated non-terrorists), a burden borne exclusively by Muslims. Certainly, society should have the opportunity to address this sort of disparity.

All else equal, society is (or should be) more willing to accept mistaken action burdens that are uniformly distributed than mistaken action burdens inflicted only on a sub-group. But the assessment might depend on how the burdened sub-group is defined. If, as in our example, the sub-group is defined by essentially immutable features – and especially if it is defined by suspect features such as race or religion – we should be particularly concerned about bias and disparate impact. For proxies employing such features, gaming costs are high and, if they are used at

[33] This discussion, for example, arose in the confirmation hearings of Brett Kavanaugh, where supporters of the now Supreme Court Justice argued that, if he was not found guilty at a criminal court, he should be confirmed. While this argument has some value, such supporters missed that the cost of false positives and false negatives are enormously different for putting someone in prison and for confirming them for the Supreme Court. Arguably, society would not want someone with, for example, 50% of having sexually assaulted women to be put in prison, and society would also not want someone with 50% of having sexually assaulted women to be in the Supreme Court.

all, the proxy's reliance on such features should be publicized so that there can be critique and debate. Disclosing the proxy also affords the opportunity for contestation as to the claim that focusing on a particular group is, in fact, producing a low false negative rate.

Even when a proxy defines a disfavored sub-group by a feature or set of features that appear innocuous, there still seems to be something worrisome about imposing unequal burdens on some members of society based on a seemingly innocuous characteristic with no apparent relationship to the decision at hand. We might also worry that the proxy we are employing is serving as, well, a proxy, for a more fundamental characteristic that the algorithm is representing by some complicated function of many seemingly innocuous traits. By employing a secret proxy in such cases, society loses the potential for public scrutiny that might help to uncover such problems.

C. Error Types and Disclosure: Summing Up

The consideration of whether to keep a decision-making algorithm opaque for fear of "gaming the system" should begin by asking whether the proxy meets the basic prerequisites for gaming. Sometimes it will be clear from this initial inquiry that at least some aspects of the proxy or algorithm can be disclosed without creating serious gaming issues.[34]

Beyond those prerequisites, it is crucial to understand what types of errors are produced by the proxy. Some proxies are uncontroversially valuable, in that they have low rates of mistaken action and inaction outcomes. Such proxies are likely to correlate with the characteristics that would ideally drive decision-making and thus unlikely to be easily gameable. Highly accurate proxies thus can, and should, be publicly disclosed, except in certain cases in which the proxy is gameable due to its temporal or incidental connection to decision characteristics of interest. Another group of proxies are uncontroversially of low value, in that they produce high rates of both mistaken actions and mistaken inaction outcomes. Proxies of this sort are often gameable in principle, but disclosure also may permit beneficial error correction and compliance that might outweigh the costs of socially undesirable gaming. Because disclosure of highly inaccurate proxies also allows for critique of the decision-maker's choice to deploy such an inaccurate proxy, its benefits will often (perhaps ordinarily) outweigh the potential risks of socially undesirable gaming as well.

[34] See Cofone & Strandburg, supra note X

The question of disclosure is more complicated when there is an imbalance between rates of mistaken action and inaction. In these situations, the main cause for concern is the possibility that the asymmetry arises because the proxy does not apply even-handedly to different sub-groups of the population. Drug investigation provides a real-life example of this concern with asymmetric error rates. While studies suggest that drug use is similarly prevalent among individuals of various races,[35] enforcement is heavily skewed toward African Americans, meaning that the rate of mistaken inaction regarding other racial groups is high. Some investigation tactics, such as "stop-and-frisk" apparently have high rates mistaken action as well, making them both discriminatory and of limited use.[36] But even if we consider a proxy with a low rate of mistaken action, we should be disturbed if it produces a high rate of mistaken inaction because of disparate impact on socially disfavored groups. In that case, we should consider combining it with other proxies, or replacing it, to reduce the disparity. Public disclosure can help to inform that decision.

It is also possible, however, that disclosure can facilitate gaming within a favored sub-group. If there is reason to think that bias is unlikely, this may be a situation where the benefits of opacity outweigh the benefits of disclosure. But it is best to be cautious in coming to that tempting conclusion. Rather than accepting that opacity is necessary to avoid gaming, it may often be better to insist that decision-makers employ more accurate proxies, which can in turn be disclosed.

VII. Conclusion

We build on the literature on algorithmic transparency that shows that disclosure is often of high social value, and the literature on algorithmic gaming that points to an important risk of disclosure: that subjects may game the system. We highlight that principal-agent problems are common in these games: decision-makers may shirk or have a divergence between their private interest and the public interest. We propose that an algorithm's performance from a social perspective is determined by its accuracy (noisiness of its proxies), the balance between tradeoffs among mistaken action and inaction outcomes, and the extent to which the algorithm is realistically gameable. We anticipate that, even in situations where gaming is possible, it may often be less socially costly than algorithmic secrecy. Therefore, secrecy should not be the default policy choice.

---

[35] [add source]

[36] [source]

More specifically, disclosure is usually socially desirable for noisy proxies with high rates of mistaken action and inaction because the reduction in an already low accuracy is of low social concern. Disclosure is usually undesirable for the opposite case (proxies with low rates of mistaken action and inaction) even when these proxies can factually be gamed, but argue that these proxies are ordinarily ungameable unless they are of the temporary or incidental sort associated with particularity. For proxies with high rates of mistaken action, but low rates of mistaken inaction, or the converse, the social desirableness of disclosure depends on the tradeoff between accountability and accuracy, as well as on the extent to which the asymmetry is a consequence of the proxy's bias or disparate impact.